DOCUMENT RESUME

ED 178 597 TM 009 804

AUTHOR	Dunivant, Noel
TITLE	Procedures for Determining the Equivalence of Measures.
INSTITUTION	New York Univ., N.Y. Dept. of Psychology.
SPONS AGENCY	National Inst. of Mental Health (DHEW), Rockville, Md.
PUB DATE	Apr 79
GRANT	1-R03-MH32019-01
NOTE	40p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Analysis of Variance; *Comparative Statistics; Evaluation Methods; Goodness of Fit; Hypothesis Testing; *Mathematical Formulas; Standard Error of Measurement; Test Reliability; Tests; *Tests of Significance
IDENTIFIERS	*Test Equivalence

ABSTRACT

Eight different methods are reviewed for determining whether two or more tests are equivalent measures. These methods vary in restrictiveness from the Wilks-Votaw test of compound symmetry (which requires that all means, variances, and covariances are equal), to Joreskog's theory of congeneric tests (which requires only that the tests are measures of the same attribute except for errors of measurement). A formal statement of the problems is given, and the various definitions or criteria of equivalence are discussed. Notation conventions are presented. Then each of the eight statistical tests for evaluating the hypothesis of equivalence is described and compared with the others. These procedures are written in a common algebra and their Type I error rates and power are compared, as are the properties of their estimators and their sensitivity to violations of their assumptions. A numerical example is worked to illustrate the application of the techniques to an educational measurement problem. The eight tests are: Wilks-Votaw; McNemar's; Forsyth-Feldt-McNemar; Forsyth-Feldt-Normal Curve; Lord's; Lord-Villegas; Kristof's; and Joreskog's. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED178597

Procedures for Determining the Equivalence of Measures

Noel Dunivant
Psychology Department
New York University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Noel Dunivant

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM009 804

Presented at the Annual Meeting of the American Educational Research Association, San Francisco, California, April 8-12, 1979. Requests for copies of this paper should be addressed to the author, Psychology Department, New York University, 6 Washington Place, 7th floor, New York, New York 10003. This research was supported by grant 1 R03 MH32019-01 from the National Institute of Mental Health.

Abstract

This paper reviews the issues and methods involved in determining if two or more tests are equivalent, that is, measure the same attribute. Eight currently available procedures for evaluating the equivalence of measures are analytically compared, including those of Forsyth and Feldt (1969, 1970), Jöreskog (1971), Kristof (1973), Lord (1957, 1973), McNemar (1958), Votaw (1948), and Wilks (1946). The comparison indicates that the methods vary widely in the restrictiveness of their assumptions and in their relative efficiencies. Where possible the circumstances under which one method is superior to the others are identified. Implications and directions for future research are discussed.

Contents

		Page
I.	Introduction	1
II.	Statement of the Problem	1
III.	Definition of Equivalence and Congeneric Tests	3
IV.	Data Organization and Notation	7
V.	Review of Methods for Determining Equivalence	8
	A. Wilks-Votaw Test	9
	B. McNemar's Test	11
	C. Forsyth-Feldt-McNemar Test	12
	D. Forsyth-Feldt-Normal Curve Test	13
	E. Lord's Test	14
	F. Lord-Villegas Test	16
	G. Kristof's Test	17
	H. Jöreskog's Test	19
VI.	Numerical Example	23
VII.	Discussion	25
VIII.	References	26
IX.	Footnotes	29
X.	Tables	30
XI.	Figures	35

Procedures for Determining the Equivalence of Measures

Introduction

This review is concerned with the problem of determining if two tests or assessment procedures measure the same underlying attribute. When two tests measure the same trait except for errors of measurement their correlation corrected for attenuation is 1.0, and they are said to be equivalent measures. Equivalent tests may differ in means and units of measurement. Statistically, the problem of demonstrating that two tests are equivalent is the same as establishing that a perfect linear relation exists between the true scores of the two tests. Ideas about equivalent measuring instruments have played a major role in the development of reliability formulations in classical test theory (Gulliksen, 1950; Lord & Novick, 1968). In recent years procedures for construct validation have emphasized the importance of determining if two different operational definitions measure the same construct (Campbell & Fiske, 1959; Cronbach, 1971; Cronbach & Meehl, 1955). The principle of equivalent measures or multiple indicators has become a major component of model building and theory testing in psychology, education, sociology and economics over the past decade (Blalock, 1969; Jöreskog & Goldberger, 1975). Additionally, there has been some interest in testing whether two variables are perfectly or "functionally" related (Isaac, 1970; Werts, Linn, & Jöreskog, 1973). In the field of developmental psychology several researchers have been concerned with the question of whether a test measures the same psychological process at a later age that it measured at an earlier age (Nesselroade & Reese, 1973). Thus, in many fields of education and the social sciences a concern has arisen with the problem of determining if two sets of measurements are equivalent.

Partially as cause and partially as effect of this interest statisticians have proposed five new procedures since 1969 for testing the hypothesis that true scores are perfectly correlated. There were only three such tests available in 1968 when Gulliksen surveyed the methods for determining the equivalence of measures. Today researchers have to choose from no fewer than eight different statistical methods when they wish to test if two assessment procedures measure the same latent trait. In general these tests are based on different sets of assumptions and are derived from different hypothesis testing and estimation techniques, e.g., method of moments and maximum likelihood. Unfortunately, most of these eight tests have not been compared either analytically or empirically, so that interested investigators have no basis for preferring one over the other methods. The methods probably do not perform equally well, however. Researchers wish to use the test with optimal properties for their particular problem and are, therefore, confronted with the problem of choosing the "best" method in the face of inadequate information.

The purposes of this review are to describe each of these procedures, to write them in a common algebra, to compare them in terms of the hypotheses they test, their Type I error rates, their power, the properties of their estimators, and their sensitivity to assumption violations. In the next section a formal statement of the problem is given and then I review various definitions or criteria of equivalence which have been proposed. Following that, notation conventions are presented. Then each of eight statistical tests for evaluating the hypothesis of equivalence is described and compared with the others. A numerical example is worked to illustrate the application of the techniques to an educational measurement problem Finally, conclusions are drawn and directions for future research are discussed.

Statement of the Problem

The problem to be studied under this proposal involves the performance of significance tests of the hypothesis that two tests measure the same attribute except for errors of measurement. For the purpose of exposition let us assume that the following simple measurement model holds for two observed tests, x and y:

$$x = X + e_x, \text{ and} \tag{1}$$

$$y = Y + e_y \tag{2}$$

where x and y are observed scores, X and Y are unobserved true scores or latent variables, and e_x and e_y are random errors of measurement. By definition x and y are equivalent measures if the correlation between the true scores X and Y is unity.

The correlation among the latent variables can be estimated in several ways. If the reliabilities of the tests are known, then the correlation between the true scores (r_{YX}) can be estimated by applying Spearman's (1904a, 1904b, 1907, 1910) correction for attenuation to the sample correlation between the observed variables ($\hat{r}_{yx}$):

$$\hat{r}_{YX} = \frac{\hat{r}_{yx}}{\sqrt{(\hat{r}_{yy})\ (\hat{r}_{xx})}} \tag{3}$$

If there are multiple measures or indicators of X (say, x_1 and x_2) and Y (say y_3 and y_4), alternative estimators of r_{YX} are available. For example, Lord[2] (1957) gives the maximum likelihood estimator as

$$\hat{r}_{YX} = \frac{\hat{\sigma}_{13} + \hat{\sigma}_{23} + \hat{\sigma}_{14} + \hat{\sigma}_{24}}{4\sqrt{\hat{\sigma}_{12}\ \hat{\sigma}_{34}}} . \tag{4}$$

(For alternative formulas based on the same covariances, see Kelley [1947] Werts & Linn [1972].) Given a particular estimate of r_{XY}, the problem is to test the hypothesis that in the population $r_{YX} = 1.0$. The null and restricted alternative hypotheses may be written:

$$H_0 : r_{YX} = 1.0, \tag{5}$$

$$H_1 : r_{YX} < 1.0. \tag{6}$$

It is also possible to express these hypotheses as tests of restrictions placed on the linear model relating Y and X. Recognizing that Y and X may differ in their means and their scaling or units of measurements, we write

$$g_1 Y = c + g_2 X \tag{7}$$

where g_1 and g_2 are scale coefficients and c is a constant which is a function of the differences in means between Y and X. Since g_1 and g_2 can be absorbed into a new coefficient $g = g_1/g_2$ and a new intercept defined as $a = c/g_1$, Equation 7 expresses Y as a linear transformation of X. Note that the structural model given by Equation 7 does <u>not</u> include a stochastic term to incorporate the effects of chance disturbances or misspecification into the model. As stated, it holds that a perfect linear relation exists between Y and X. This is sometimes referred to as a functional relation in the statistical literature (Isaac, 1970; Kendall & Stuart, 1961). The null and alternative hypotheses to be tested are:

$$H_0: g_1 Y - g_2 X - c = 0, \tag{8}$$

$$H_1: g_1 Y - g_2 X - c \neq 0. \tag{9}$$

It should be apparent that since correlations are invariant to differences of scale and origin, the null hypotheses given by Equations 5 and 8 are the same.

A variety of different ways of testing Equations 5 and 8 under various types of assumptions have been proposed. In order to explicate these methods and their assumptions we now define equivalence and present Jöreskog's model of congeneric tests.

Definition of Equivalence and Congeneric Tests

In the development of classical test theory the concept of strictly parallel tests has played a crucial role. Strictly parallel tests by definition are tests which have equal means, equal variances, equal covariances and equal validities with respect to any criterion. Parallel tests also haveequal standard errors of measurement and equal reliabilities. A person has the same true score on parallel tests (Gulliksen, 1950; Lord & Novick, 1968). Gulliksen (1968) argues that scores on parallel tests are "interchangeable." That is, scores from one parallel can be substituted for those for another parallel test without any loss of information whatsoever. Obviously, parallel tests satisfy the criterion for equivalence, that is, they measure the same trait except for errors of measurement. However, other tests, which do not meet the rigorous requirements for parallelism, likewise satisfy the criterion. For example, tau equivalent tests, i.e., tests which meet all of the requirements for parallelism except that they have different standard errors and consequently unequal reliabilities (Lord & Novick, 1968), measure the same underlying attribute. While tau equivalent tests measure the same construct, they do so with varying degrees of accuracy (reliability). The model for essentially tau equivalent tests goes even further by relaxing the restriction on equal means. Thus, an individual will not necessarily have equal true scores on essentially tau equivalent tests; however, all true scores on one essentially tau equivalent test will differ from those on another essentially tau equivalent test only by a constant.

The most general model for tests that measure the same attribute except for errors of measurement is Jöreskog's (1971) theory of congeneric tests. In this model almost all of the restrictions on parallel tests have been eliminated. The tests need not possess equal true means, equal true or error variances, or equal reliabilities. Individuals do not have equal scores on congeneric tests. This implies that congeneric tests have different origins or means and different scales or units of measurement. However, true scores on one congeneric test are a perfect linear function of true scores on another congeneric test. Congeneric tests meet Gulliksen's (1968) criteria for "scientific equivalence," that is, they measure the same underlying attribute. In factor analytic terminology congeneric tests reflect a single general factor. While congeneric tests are not strictly interchangeable or substitutable and are not of equal accuracy, they contain information about the same latent variable which underlies each of the tests. A detailed comparison of the various test models is provided in Table 5. Now these ideas about the equivalence of measures is given a more precise mathematical form by developing Jöreskog's (1970, 1971, 1974) theory of congeneric tests.

Let us assume that there are two replicate measures on each of two tests x and y. By definition x_1 is congeneric with x_2 and y_3 is congeneric with y_4. While the replications may represent two alternate forms or test-retest measurements on the same test, in most cases these replications are obtained by spliting tests x and y into halves. Such split halves may or may not be parallel, but they must be congeneric for the following developments to hold. These congeneric replicates or multiple indicators are necessary in order to identify the parameters of the true and error distributions. Without the additional information provided by the multiple measurements the true score correlation cannot be estimated or tested against a hypothesized value (e.g., 1.0). The reader is referred to the text by Hanushek and Jackson (1977) for an excellent introduction to the problem of identification.

According to classical theory the observed scores can be written as a linear composite of true and error components:

$$x_1 = X_1 + e_1, \quad (10)$$
$$x_2 = X_2 + e_2, \quad (11)$$
$$y_3 = Y_3 + e_3, \quad (12)$$

$$y_4 = Y_4 + e_4, \tag{13}$$

where X_1, X_2, Y_3 and Y_4 are true scores and the e_i are random errors of measurement. Under the classical assumptions true scores are not correlated with errors, and errors are all mutually uncorrelated. Since by definition the correlation between congeneric tests is 1.0, $r_{X_1X_2} = r_{Y_1Y_2} = 1.0$. Therefore, new random variables X and Y can be defined which are perfectly linearly related to the true scores on the individual tests X_1, X_2 and Y_3, Y_4, respectively, as follows:

$$X_1 = m_1 + b_1X, \tag{14}$$
$$X_2 = m_2 + b_2X, \tag{15}$$
$$Y_3 = m_3 + b_3Y, \tag{16}$$
$$Y_4 = m_4 + b_4Y. \tag{17}$$

Substitution of Equations 14 through 17 into Equations 10 through 13 yields the congeneric measurement model:

$$x_1 = m_1 + b_1X + e_1, \tag{18}$$
$$x_2 = m_2 + b_2X + e_2, \tag{19}$$
$$y_3 = m_3 + b_3Y + e_3, \tag{20}$$
$$y_4 = m_4 + b_4Y + e_4. \tag{21}$$

In matrix form this result may be written

$$\begin{pmatrix} x_1 \\ x_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \end{pmatrix} + \begin{pmatrix} b_1 & 0 \\ b_2 & 0 \\ 0 & b_3 \\ 0 & b_4 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}$$

$$\underline{x} = \underline{m} + \underline{B}\ \underline{t} + \underline{e}\ . \tag{22}$$

where $\underline{x}$ is a vector of observed scores for an individual, $\underline{m}$ is a vector of means, $\underline{B}$ is a matrix of scaling coefficients, $\underline{t}$ is a vector of true scores, and $\underline{e}$ is a vector of errors of measurement.

Without loss of generality we take the true scores to be expressed in standardized form ($\mathcal{E}(X) = \mathcal{E}(Y) = 0$, VAR (X) = VAR (Y) = 1.0).

Then the structural model relating Y to X is

$$g_1Y = g_2X, \text{ or alternatively} \tag{23}$$

$$Y = \frac{g_1}{g_2}X = gX, \text{ and} \tag{24}$$

the covariance matrix of the vector $\underline{t}$ is

$$\Sigma_{XY} = \begin{pmatrix} 1.0 & r_{XY} \\ r_{YX} & 1.0 \end{pmatrix}\ . \tag{25}$$

From Jöreskog's (1970) covariance structure analysis the covariance matrix of observed variables can be written as a function of the parameter matrices

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{pmatrix} = \begin{pmatrix} b_1 & 0 \\ b_2 & 0 \\ 0 & b_3 \\ 0 & b_4 \end{pmatrix} \begin{pmatrix} 1.0 & r_{XY} \\ r_{YX} & 1.0 \end{pmatrix} \begin{pmatrix} b_1 & b_2 & 0 & 0 \\ 0 & 0 & b_1 & b_2 \end{pmatrix}$$

$$+ \begin{pmatrix} \sigma_{e_1}^2 & 0 & 0 & 0 \\ 0 & \sigma_{e_2}^2 & 0 & 0 \\ 0 & 0 & \sigma_{e_3}^2 & 0 \\ 0 & 0 & 0 & \sigma_{e_4}^2 \end{pmatrix}, \text{ or}$$

$$\Sigma_{xy} = \underline{B}\ \Sigma_{XY}\ \underline{B}' + \Sigma_{ee}. \tag{18}$$

It is also possible to specify the parametric structure of the mean vector within Joreskog's (1970) Analysis of COVariance Structure model (ACOVS). Let $\underline{D}$ be an N x 4 matrix of scores on tests x_1, x_2, y_3, and y_4 from a sample of size N, $\underline{E}$ an N x 1 matrix of ones, and $\underline{G}$ a 4 x 4 matrix specifying constraints on the mean vector. Then the population means are structured as

$$\mathcal{E}(\underline{D}) = \underline{E}\ \underline{m}'\ \underline{G}. \tag{26}$$

When $\underline{G}$ is the identity matrix, $\mathcal{E}(\underline{D}) = \underline{m}'$.

If x and y are congeneric, then r_{XY} will equal 1.0. That is $g_2/g_1 = g = r_{XY} = 1.0$. Thus, testing the hypothesis that two tests x and y measure the same attribute except for errors of measurement (or that x and y are scientifically equivalent or congeneric) reduces to testing the hypothesis that $r_{XY} = 1.0$. Under congeneric assumptions the parameters b_1, b_2, b_3, b_4, m_1, m_2, m_3, m_4, $\sigma_{e_1}^2$, $\sigma_{e_2}^2$, $\sigma_{e_3}^2$, $\sigma_{e_4}^2$ are free to assume any values whatsoever. A path diagram for the congeneric model is presented in Figure 1. Since the true scores are standardized, the coefficient g must equal

---

Insert Figure 1 about here

1.0 when x and y are congeneric.

As was indicated above parallel, tau equivalent and essentially tau equivalent tests are special cases of the general congeneric model. It is easy to illustrate the restrictions which must be imposed on the general congeneric model in order to obtain these special cases. For example, if the two halves of x and those of y are to be essentially tau equivalent, then b_1 must equal b_2 and b_3 must equal b_4. No restrictions are placed on the means or error variances under this model. However, since the half tests are scaled in the same units of measurement, scores can be directly added or subtracted, e.g., $x = x_1 + x_2$. (In order for scores on congeneric tests to be combined, scores from one test must first be converted into the unit of measurement of the other.) If we further require that the half tests be tau equivalent, then restrictions on the means must be satified such that $m_1 = m_2$ and $m_3 = m_4$. Such restrictions may be specified by choosing $\underline{G}$ to be:

$$\begin{pmatrix} 0.5 & 0.5 & 0.0 & 0.0 \\ 0.5 & 0.5 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.5 & 0.5 \\ 0.0 & 0.0 & 0.5 & 0.5 \end{pmatrix} . \tag{27}$$

Then the vector of expected values for any person sampled from the population will be

$$\mathcal{E}(\underline{x}') = \left(\frac{m_1 + m_2}{2} , \frac{m_1 + m_2}{2} , \frac{m_3 + m_4}{2} , \frac{m_3 + m_4}{2} \right). \tag{28}$$

Suppose that x_1 and x_2 meet the criteria for parallelism and that y_3 and y_4 do also. In this case the following restrictions must hold:

$$b_1 = b_2,\ b_3 = b_4;\ m_1 = m_2,\ m_3 = m_4;\ \sigma_{e_1} = \sigma_{e_2} ,\ \sigma_{e_3} = \sigma_{e_4} . \tag{29}$$

Under these specificiations scores from x_1 are completely interchangeable with those on x_2; also, y_3 and y_4 scores may be freely substituted.

Thus far we have considered the questions of the equivalence of x_1 and x_2 and the equivalence of y_3 and y_4. Now let us examine the additional restrictions on the congeneric model imposed by test score models for the equivalence of x and y. It has already been noted that r_{XY} must be unity for x and y (or more properly x_1, x_2, y_3 and y_4) to be congeneric. In order for the four tests to satisfy the assumptions of essentially tau equivalent tests, all four tests must be scaled in the same units, i.e.,

$$b_1 = b_2 = b_3 = b_4 \tag{30}$$

If the tests are tau equivalent, then the parametric structure of the mean vector must conform to a $\underline{G}$ matrix of the form

$$\begin{pmatrix} .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \end{pmatrix} , \tag{31}$$

which produces a vector of means which are all equal:

$$\mathcal{E}(\underline{x}') = \left(\frac{m_1 + m_2 + m_3 + m_4}{4} , \frac{m_1+m_2+m_3+m_4}{4} , \frac{m_1+m_2+m_3+m_4}{4} , \frac{m_1+m_2+m_3+m_4}{4} \right). \tag{32}$$

Of course, the condition given by Equation 32 is true only when

$$m_1 = m_2 = m_3 = m_4 . \tag{33}$$

Taken together Equations 30 and 33 define the model of tau equivalent tests.

Now suppose that we impose the most stringent restrictions on the general model; that is, that the four tests are completely interchangeable or parallel. Then in addition to the conditions given by Equations 30 and 33, the restriction that the tests have equal standard errors must apply:

$$\sigma_{e_1} = \sigma_{e_2} = \sigma_{e_3} = \sigma_{e_4} . \tag{34}$$

Since parallel tests have equal means, equal units of measurement, and equal standard errors, the parametric structures of the means, variances and covariances are completely specified by three values; m_1, b_1, and σ_{e_1}. It was pointed out earlier that parallel tests are equally reliable, where reliability is given its traditional definition as the ratio of true score variance to the sum of the true score variance and the error variance. Referring to Equations 18 through 21 we observe that the scaling coefficients b_j are defined as ordinary regression coefficents:

$$b_j = \frac{\text{Cov}(x_j\ X)}{\text{Var}(X)} = \text{Cov}(x_j\ X), \tag{35}$$

since Var (X) = 1.0. It follows that b_j^2 is the true score variance of test j. Thus, the reliability of test j is

$$r_{jj} = \frac{b_j^2}{b_j^2 + \sigma_{e_j}^2} . \tag{36}$$

Obviously, for all parallel tests, i.e., tests which satisfy Equations 30, 33, and 34, the reliabilities will be equal.

Before concluding this section a final point should be made concerning the rank of the matrix Σ_{xy}. If x_1, x_2, y_3 and y_4 are all congeneric, then Equation 22 can be rewritten as

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} t + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}$$

$$\underline{x} = \underline{m} + \underline{b}t + \underline{e}, \tag{37}$$

and Equation 22 as

$$\Sigma_{xy} = \underline{b}\,\underline{b}' + \Sigma_{ee}, \tag{38}$$

which is formally equivalent to a factor analytic model with one common factor. When r_{XY} is unity, the rank of Σ_{xy} and Σ_{XY} equals one. Testing the restriction that $r_{XY} = 1.0$ is equivalent to testing whether a single factor model fits the data (Gulliksen, 1968; Joreskog, 1974). The path diagram for the one factor model is depicted in Figure 2.

Insert Figure 2 about here

Data Organization and Notation

In order to facilitate our comparison of the several statistical procedures for determining the equivalence of measures, a common data layout and notation will be employed for all the procedures. First, we assume that there are measurements on two tests x and y which have been split into congeneric halves: x_1 and x_2, and y_3 and y_4 as was described above. Thus, the congeneric measurement model specified in Equations 18 through 21 holds. The scores for N persons on the four tests are organized according to the schema presented in Table 1, which also illustrates the notational conventions that have been adopted.

Insert Table 1 about here

The population covariance matrix of the vector (x_1, x_2, x, y_3, y_4, y) may be expressed in lower triangular form as

$$\Sigma_{(x_1,x_2,x,y_3,y_4,y)} = \begin{pmatrix} \sigma_{11} & & & & & \\ \sigma_{21} & \sigma_{22} & & & & \\ \sigma_{x1} & \sigma_{x2} & \sigma_{xx} & & & \\ \sigma_{31} & \sigma_{32} & \sigma_{3x} & \sigma_{33} & & \\ \sigma_{41} & \sigma_{42} & \sigma_{4x} & \sigma_{43} & \sigma_{44} & \\ \sigma_{y1} & \sigma_{y2} & \sigma_{yx} & \sigma_{y3} & \sigma_{y4} & \sigma_{yy} \end{pmatrix} . \tag{39}$$

The estimator of Σ derived from a sample of size N will be denoted $\hat{\Sigma}$, the elements of which are deviation sums of squares and cross products devided by N-1 and symbolized as $\hat{\sigma}$. The estimator of the population correlation matrix is derived from $\hat{\Sigma}$ and may be written

$$\underset{(x_1,x_2,x,y_3,y_4,y)}{\underline{R}} = \begin{pmatrix} 1.0 & & & & & \\ r_{21} & 1.0 & & & & \\ r_{x1} & r_{x2} & 1.0 & & & \\ r_{31} & r_{32} & r_{3x} & 1.0 & & \\ r_{41} & r_{42} & r_{4x} & r_{43} & 1.0 & \\ r_{y1} & r_{y2} & r_{yx} & r_{y3} & r_{y4} & 1.0 \end{pmatrix}. \qquad (40)$$

We will refer to the entries in Σ $(x_1, x_2, x, y_3, y_4, y)$ and $\hat{\underline{R}}$ $(x_1, x_2, x, y_3, y_4, y)$ frequently in the sections to follow. The definitions of the other vectors and matrices remain as given in the preceding sections, e.g., $\underline{m}$, $\underline{B}$ Σ_{XY}, etc.

Review of Methods for Determining Equivalence

The purpose of this section is to explicate and compare eight statistical methods designed for determining if two tests x and y measure the same underlying attribute. We will review the tests of Forsyth and Feldt (1969, 1970), Jöreskog (1971), Kristof (1973), Lord (1957, 1973), McNemar (1958), and Votaw (1948) and Wilks (1946). These procedures will be presented in an order which roughly corresponds to the restrictiveness of the assumptions on which the tests are based. Our review will begin with the Wilks-Votaw tests which use the most restrictive models and conclude with Jöreskog's method, which is the most general. To the extent possible I have tried to follow the same outline in describing each of the methods. At the outset the statistical model and its assumptions are stated. Then the null and alternative hypotheses which are tested by the procedure are specified. Next we provide the computational formulas for the test statistic and describe how its significance is evaluated. If provided by the test developer, the estimator of r_{XY} under a true alternative hypothesis is presented. Finally, an evaluation of the test is made, For example, evidence which contradicts the validity of the test is discussed. If the efficiency of a test relative to one or more other tests is known, the superiority of the method is pointed out. We also demonstrate that some of the tests differ only in computational methods, e.g., in the way the likelihood function is evaluated. They are identifical statistical tests in all other respects.

In an attempt to clarify the differences between the eight methods a table has been prepared in which the assumptions and null hypotheses of each of the procedures are identified. Table 2 identifies most of the possible restrictions which can be placed on the general structural model relating the observed, true and error components of four tests x_1, x_2, y_3 and y_4. For example, assumptions concerning the multivariate normality of the variables, the parallelism of x_1 and x_2, etc. are listed.

Insert Table 2 about here

The reader will observe that three different symbols are used to designate which assumptions, hypotheses and properties are appropriate for the methods. The letter "Y" (Yes) indicates that a given procedure requires the stated assumption. If the procedure incorporates the assumption into the null hypothesis to be tested, then it is designated with the letter "H" (Hypothesis). The difference between Y and H assumptions can be easily illustrated. The formulation of each statistical test may be stated as: "given that 'Y' assumptions are true, test the verity of the 'H' assumptions." Lastly, the letter "N"(No) is used to identify those assumptions which are not required for the procedure to be valid. Frequent reference will be made to Table 2 throughout the following review.

Wilks-Votaw Test

The first tests which we consider were developed by Wilks (1946) and Votaw (1948) to determine if two or more measures are interchangeable, that is, if they have equal means, equal variances and equal covariances. Of course, this is a much more restrictive test than that the measures are only scientifically equivalent. However, there are occasions in which a very strict equivalency or interchangeability will be of interest. Also some of the methods to be considered later require that the half tests be parallel. The Wilks-Votaw statistical tests can be used to determine if the assumption of parallelism is tenable for these procedures.

Wilks (1946) provided test statistics for the hypotheses that 1) all means, all variances, and all covariances are equal, and 2) all variances and all covariances are equal. These were extended by Votaw (1948) to include the hypothesis of equal validities (covariances) with a set of criterion variables, a condition resulting in compound symmetry of the covariance matrix. Our treatment will be concerned exclusively with Wilks' tests and will follow Gulliksen (1950) closely. First, we consider the assumptions which are required for the tests to be valid. The most important assumptions as indicated in Table 2 are that the data are based on a large sample, that the four tests have a joint multivariate normal distribution, and th the standard test theory assumptions hold, e.g., true and error scores uncorrelated and errors mutually uncorrelated with expectations zero.

Now we consider Wilk's likelihood ratio test of the null hypothesis of equal mean equal variances and equal covariances for the four tests x_1, x_2, y_3, and y_4:

$$H_0: \quad m_1 = m_2 = m_3 = m_4 \tag{41}$$

$$b_1 = b_2 = b_3 = b_4 \tag{42}$$

$$\sigma_{e_1} = \sigma_{e_2} = \sigma_{e_3} = \sigma_{e_4} \tag{43}$$

$$r_{XY} = 1.0 \tag{44}$$

This hypothesis is tested against the alternative that at least one of the four conditions specified by Equations 41, 42, 43 and 44 does not hold. Note that this null may be false even the correlation between true scores X and Y is perfect.

Wilk's test statistic L_{mvc} (means, variances, covariances) equals:

$$L_{mvc} = \frac{\left|\hat{\Sigma}_{xy}\right|}{\hat{\bar{\sigma}}^2 \left[1 + 3\hat{\bar{r}}\right] \left[\hat{\bar{\sigma}}^2 \left(1 - \hat{\bar{r}}\right) + \hat{\sigma}_m^2\right]^3}, \tag{45}$$

$\left|\hat{\Sigma}_{xy}\right|$ is the determinant of the 4x4 covariance matrix of observed scores x_1, x_2, y_3 and y_4,

$$\hat{\bar{\sigma}} = \frac{\sum_{j=1}^{4} \hat{\sigma}_{jj}}{4}, \text{ estimate of average variance,} \tag{46}$$

$$\hat{\bar{r}} = \frac{\sum_{j \neq k=1}^{4} \hat{\sigma}_{jk}}{12\,\hat{\bar{\sigma}}}, \text{ estimate of average correlation,} \tag{47}$$

$$\hat{\sigma}_m^2 = \frac{(\bar{x}_{.1} - M)^2 + (\bar{x}_{.2} - M)^2 + (\bar{y}_{.3} - M) + (\bar{y}_{.4} - M)}{3}, \text{ estimate of variance of the means, where} \tag{48}$$

$$M = \frac{\bar{x}_{.1} + \bar{x}_{.2} + \bar{y}_{.3} + \bar{y}_{.4}}{4}, \text{ mean of test means.} \tag{49}$$

For small samples ($N \leq 100$), computed values of L_{mvc} can be referred to the table presented in Wilks' (1946) article. In large sample research, the quantity-N ln L_{mvc} is approximately distributed as chi sequare with 11 degrees of freedom for our problem when H_0 is true. Thus, if $\chi_{mvc}^2 = - N \ln L_{mvc}$ does <u>not</u> exceed the tabled value of chi square for 11 degrees of freedom, the investigator <u>cannot</u> reject the hypothesis of interchangeabilty. Under a true null hypothesis,
the best indicator (designated with zero subscripts) of the three parameters of the distributions of the observed scores according to Wilks are

$$M_0 = M, \tag{50}$$

$$\hat{\sigma}_0^2 = \bar{\hat{\sigma}}^2 + \frac{3\hat{\sigma}_m^2}{4}, \tag{51}$$

$$\hat{r}_0 = \frac{\bar{\hat{\sigma}}^2 \bar{\hat{r}} - \frac{\hat{\sigma}_m^2}{3}}{\hat{\sigma}_0^2} \tag{52}$$

If the sample value of χ_{mvc}^2 exceeds the critical chi square value, then the equality of only the variances and covariances can be tested. The new null hypothesis becomes

$$H_0: \quad b_1 = b_2 = b_3 = b_4, \tag{53}$$

$$\sigma_{e_1} = \sigma_{e_2} = \sigma_{e_3} = \sigma_{e_4}, \tag{54}$$

$$r_{XY} = 1.0 \tag{55}$$

which is tested against the alternative that either Equation 53, 54 or 55 is untrue. The test statistic can be given in terms of quantities already defined:

$$L_{vc} = \frac{\left|\hat{\Sigma}_{xy}\right|}{\bar{\hat{\sigma}}^2 [1 + 3\bar{\hat{r}}] [\bar{\hat{\sigma}}^2 (1 - \bar{\hat{r}})]^3}. \tag{56}$$

Again, the quantity- N ln L_{vc} is approximated distributed as chi square in large samples with eight degrees of freedom when the null is true. If the researcher fails to reject H_0, the optimal estimators are:

$$M_{0j} = \bar{x}_{.1}, \bar{x}_{2.}, \bar{y}_{3.}, \bar{y}_{4.}, \tag{57}$$

$$\hat{\sigma}_0^2 = \bar{\hat{\sigma}}^2, \tag{58}$$

$$\hat{r}_0 = \bar{\hat{r}}. \tag{59}$$

In summary, if one wishes to determine not only if a set of measures are equivalent but also if they are interchangeable with one another, then the likelihood ratio tests of Wilks (1946) and Votaw (1948) can be used. Although the test statistics possess some optimal large sample properties, little is known about the adequacy of the tests in small sample research. The effects of violating various assumptions, such as, multivariate normality or uncorrelated errors, have not been investigated. Finally, we note that Joreskog's ACOVS computer program for performing maximum likelihood

estimation (Jöreskog, Gruvaeus & van Thillo, 1970) could be used to compute χ^2_{mvc} or χ^2_{vc} and the ML estimates using a different algorithm (see section on Jöreskog below).

McNemar's Test

Working within an analysis of variance framework McNemar (1958) devised a test for "whether a correlation coefficient is less than unity by an amount greater than that attributable to errors of measurement" (p. 259). The procedure is based on the observation that when thecorrelation corrected for attenuation is 1.0 (i.e., the tests measure the same attribute except for errors of measurement), the Individuals x Tests interaction in the three-way ANOVA design represented in Table 1 (Individuals x Tests x Replications) will be exactly zero. McNemar (1958) pointed out that this correspondence between attenuation and interaction leads to a relatively simple test of the hypothesis that $r_{XY} = 1.0$.

The simplicity of the method, however, is purchased at the expense of some restrictive assumptions which must be met in order for the test to be valid, i.e., for its probability value and confidence intervals to be correct as specified (Wonnacott & Wonnacott, 1972). Reference to Table 2 informs us that all of the tests have to be in a common metric in order to use the analysis of variance technique. Additionally, ANOVA requires homogeneity of variance which means that all of the tests must have equal variances. McNemar tries to achieve these conditions artificially by standardizing the scores on each test. However, with standard scores we must assume that the tests have equal error variances or, equivalently, equal reliabilities. It is most likely that this condition will be met when x and y have been developed as parallel tests. Table 2 also reveals that the McNemar test requires that all of the standard classical theory assumptions hold.

The null hypothesis which is tested by this method is $H_0 : r_{XY} = 1.0$, given that $b_1 = b_2$, $b_3 = b_4$, $\sigma_{e_1} = \sigma_{e_2}$, $\sigma_{e_3} = \sigma_{e_4}$, and $r_{xx} = r_{yy}$. The alternative is that $r_{XY} < 1.0$.

To compute the appropriate F-statistic for the null hypothesis one can perform an Individuals x Tests x Replications ANOVA on standardized scores laid out as in Table 1. If the obtained value of F for the I x T interaction (using either the proper or a pooled error term) does not exceed the critical value of F for the nominal alpha level, then the null hypothesis of $r_{XY} = 1.0$ is supported. McNemar (1958, Equation 7) derives the F-ratio (based on a pooled error term, hence the subscript 2), for the null hypothesis in terms of sample correlations between the observed tests:

$$F_2 = \frac{2 + \hat{r}_{12} + \hat{r}_{34} + \hat{r}_{13} + \hat{r}_{14} + \hat{r}_{23} + \hat{r}_{24}}{2 - \hat{r}_{12} - \hat{r}_{34}}, \qquad (60)$$

with N-1 and 2(N-1) degrees of freedom. When the probability value of F_2 is less than the stated alpha, we conclude that x and y are not equivalent or congenetic tests. There is also the possibility that the null will be rejected because x_1, x_2 and y_3, y_4 do not meet the assumptions of parallelism. Therefore, it seems judicious to use the Wilks-Votaw tests to determine if the half tests are parallel before proceeding with McNemar's test (cf. Lord, 1957).

The ease of computing McNemar's test statistic and the fact that its properties are not dependent upon sample size strongly recommend this procedure. However, Monte Carlo studies by Forsyth and Feldt (1969, 1970) have demonstrated that the method is

is very sensitive to violations of the assumption of equal reliabilities. Unequal reliabilities lead to more frequent rejection of the null hypothesis than the nominal alpha value indicates, i.e., Type I errors are not sufficiently controlled. Technically, we would say that the test is "unsafe" (Wonnecott & Wonnecott, 1972). Furthermore the disturbing effects of unequal reliabilities increase with sample size. Forsyth and Feldt (1970) proposed a modification of McNemar's test, to which we now direct our attentions.

Forsyth-Feldt-McNemar Test

Forsyth and Feldt (1970) show that when the assumption of equal reliabilities is not satisfied F_2 does not follow a central F distribution with N-1 and 2N-2 degrees of freedom. Instead the quantity F_2/K is distributed as central F under the null hypothesis ($r_{XY} = 1.0$) when the reliabilities of x and y are not equal, where

$$K = \frac{1 + \bar{r}_{rel} - 2\sqrt{r_{12}r_{34}}}{1 - \bar{r}_{rel}} \quad , \text{ and} \tag{61}$$

$$\bar{r}_{rel} = \frac{r_{xx} + r_{yy}}{2} \quad . \tag{62}$$

However, K is defined in terms of population parameters which are never known exactly in practice. However, Forsyth and Feldt (1970) suggested "that a 'good' estimate of K may lead to procedures which provide reasonably accurate control of Type I errors. An appealing estimate of K is that obtained by substituting sample estimates of the reliabilities ($\hat{r}_{12}$, $\hat{r}_{34}$, $\hat{\bar{r}}_{rel}$) into the equation"[1] (p. 200). The reliabilities needed for $\hat{K}$ could be estimated by several different methods; however, the most reasonable approach is to apply the Spearman-Brown phophecy formula (Gulliksen, 1950) to the correlations between parallel forms:

$$\hat{r}_{xx} = \frac{2\hat{r}_{12}}{1 + \hat{r}_{12}} \quad , \text{ and} \tag{63}$$

$$\hat{r}_{yy} = \frac{2\hat{r}_{34}}{1 + \hat{r}_{34}} \quad . \tag{64}$$

If the value of $F_2/\hat{K}$ exceeds the tabled value of F for N-1 and 2N-2, then the hypothesis that the correlation corrected for attenuation equals unity is rejected at the alpha level of significance.

The results of simulation experiments suggest that the test statistic $F_2/\hat{K}$ is superior to F_2 in controlling Type I errors. Forsyth and Feldt (1970) found that when the reliabilities were not equal, the average absolute deviation from the nominal .05 alpha level was .015. The discrepancy was even smaller when the reliabilities were equal. Although the differences were small, "there [was] a definite tendency for the empirical alpha levels associated with F_2/K to be greater than the nominal α" (Forsyth & Feldt, 1970, p. 204). Differences in power between F_2 and F_2/K were evaluated only for the equal reliability case. When r_{XY} equaled .95, the power of F_2 exceeded that of $F_2/\hat{K}$, but the difference was not great. The maximum difference in power between the two tests was only .024. The Monte Carlo studies also demonstrated that the Forsyth-Feldt-McNemar test gave better control of Type I errors than did an approximate test devised by Forsyth and Feldt (1969), which was based on normal curve theory. We consider this test in the next section.

Before concluding three additional points should be emphasized. First, the Forsyth-Feldt-McNemar test is based on the same assumptions as the McNemar test except for the assumption of equal reliabilities (see Table 2). Second, the exact theoretical properties of the test statistic $\underline{F}_2/\hat{K}$ are not known. For example, although $\underline{F}_2/K$ follows a central $\underline{F}$ distribution, it has not been proved that $F_2/\hat{K}$ is distributed as a central $\underline{F}$ distribution with N-1 and 2N-2 degrees of freedom under the null hypothesis. Third, the effects of assumption violations on $\underline{F}_2/\hat{K}$, e.g., nonnormal error distributions or correlated errors, have not been explored.

Forsyth-Feldt Normal Curve Test

Forsyth and Feldt (1969) investigated the empirical sampling distribution of the correlation corrected for attenuation in 72,000 simulated samples. When the population correlation between the true scores equaled 1.0, they discovered that 1) the mean r_{XY} equalled 1.0 so that the estimator behaved as if it were unbiased; 2) although the variance of r_{XY} was systematically greater than the approximate formulas of Cureton (1936) and Kelley (1947), the degree of overestimation was small for Kelley's formula; 3) the sampling distribution of r_{XY} was slightly positively skewed; and 4) the sampling distribution of r_{XY} was leptokurtic with a kurtosis index of 3.4. In spite of the peaked nature of the sampling distribution of r_{XY} Forsyth and Feldt (1970) concluded that "for suitably large samples, the sampling distribution of r_{XY} is approximately normal in form" (p. 65) and argued that standard normal curve techniques for constructing confidence intervals could be applied.

The assumptions required by Forsyth and Feldt's approximate normal curve test of the null hypothesis that the disattenuated correlation between x and y is unity are given in Table 2. In addition to the classical test theory assumptions, we see that each of the half tests must be parallel with respect to units of measurement and standard errors and that the joint distribution of x and y should be bivariate normal. Given the truthfulness of these assumptions the Forsyth-Feldt procedure tests the null against the alternative hypothesis that $r_{XY} < 1.0$.

In order to use normal curve theory to construct a significance test one must have an estimator of r_{XY}, an estimator of the standard error of r_{XY}, and a choice for the protection level against Type I errors of inference. The Forsyth-Feldt estimator of the correlation correct for attenuation is:

$$\hat{r}_{XY} = \frac{\hat{r}_{xy}}{\sqrt{\left(\frac{2\,\hat{r}_{12}}{1+\hat{r}_{12}}\right)\left(\frac{2\hat{r}_{34}}{1+\hat{r}_{34}}\right)}} , \qquad (65)$$

which is essentially the standard correction (cf. Equation 3) in which the parallel forms reliabilities have been stepped up by the Spearman-Brown formula. Under H_0 the estimator of the sampling variance of $\hat{r}_{XY}$ is defined by Forsyth and Feldt as:

$$\hat{\sigma}_{\hat{r}_{XY}}^2 = \frac{1}{4(n-2)}\left[2 + \frac{(1+\hat{r}_{12})(1+\hat{r}_{34})}{\hat{r}_{12}\,\hat{r}_{34}} + \frac{1}{\hat{r}_{12}^2} \quad \frac{1}{\hat{r}_{34}^2} - \frac{4}{\hat{r}_{12}} - \frac{4}{\hat{r}_{34}}\right] , \qquad (66)$$

which is Kelley's (1947) formula. The critical value of $\hat{r}_{XY}$ required for the rejection of the null hypothesis that x and y are scientifically equivalent measures for the α level of significance is:

$$r_{crit} = 1.0 - \underline{z}_{\alpha}\,\hat{\sigma}_{\hat{r}_{XY}}\left(\sqrt{N/(N-2)}\right) . \qquad (67)$$

The use of the multiplier $\sqrt{N/(N-2)}$ was determined "entirely on empirical grounds" (Forsyth & Feldt, 1969, p. 66). If $\hat{r}_{XY} < r_{crit}$, then the investigator rejects the hypothesis that x and y measure the same trait except for errors of measurement. Note that this test, as are all those which evaluate H_0: $r_{XY} = 1.0$, is a one-sided test since a direction restriction is placed on the alternative, i.e., H_1: $r_{XY} < 1.0$.

Monte Carlo evaluations of the Forsyth-Feldt approximate normal curve test indicated that it accorded fair control of the Type I errors. (Forsyth & Feldt, 1969, 1970). When r_{XY} equaled 1.0, the empirical alpha rate exceeded the nominal .05 level by an average of .025 (computed from entries in Table 2, Forsyth & Feldt, 1969), indicating that the null hypothesis was rejected more frequently than would be expected from the nominal alpha level. When the reliabilities of x and y were unequal (and only when they were unequal), the Forsyth-Feldt normal curve test performed better than the McNemar test (Forsyth & Feldt, 1969). However, as was alluded to in the previous section the Forsyth-Feldt approximate normal curve test did not provide as effective control of Type I errors as did the Forsyth-Feldt modification of McNemar's test in an extensive series of simulation studies (Forsyth & Feldt, 1970). Even though the Forsyth-Feldt test is based on less restrictive assumptions, the Forsyth-Feldt modification of the McNemar ANOVA method is a safer and relatively more efficient test. "The relative efficiency of two tests is the ratio of samples required to give equally accurate tests, i.e., tests with the same probability of type I and type II errors" (Wonnacott & Wonnacott, 1972, p. 416). Thus, to achieve the same degree of accuracy the Forsyth-Feldt test must be used with a larger sample than that required by the Forsyth-Feldt-McNemar test.

Lord's test

Historically the first test for the hypothesis of a perfect linear relation between true scores X and Y which did not require that the observed tests x and y be parallel, was proposed by Frederic Lord (1957) and based on the large sample chi-square approximation to the likelihood ratio statistic. Lord's (1957) procedure requires large samples for its justification and assumes that x_1 and x_2 and y_3 and y_4 meet parallelism assumptions for scaling and standard errrs, i.e., $b_1 = b_2$, $b_3 = b_4$ and $\sigma_{e_1} = \sigma_{e_2}$, $\sigma_{e_3} = \sigma_{e_4}$ (see Table 2). Furthermore, the maximum likelihood estimation technique used by Lord requires the multivariate normality of the vector (x_1, x_2, y_3, y_4). Given these assumptions, Lord (1957) defines the likelihood ratio for testing the hypothesis that $r_{XY} = 1.0$ as

$$\lambda = \frac{L(\tilde{\sigma}_{11}, \tilde{\sigma}_{12}, \ldots, \tilde{\sigma}_{44})}{L(\dot{\sigma}_{11}, \dot{\sigma}_{12}, \ldots, \dot{\sigma}_{44})}, \qquad (68)$$

where the $\tilde{\sigma}$ are ML estimates when the null hypothesis is true and $\dot{\sigma}$ are ML estimates when the alternative hypothesis is true. The quantity $-2 \ln\lambda$ is asymptotically distributed as chi square with one degree of freedom and equals

$$\chi^2 = (N-1) \ln \frac{(\tilde{\sigma}_{11} - \tilde{\sigma}_{12})(\tilde{\sigma}_{33} - \tilde{\sigma}_{34})[(\tilde{\sigma}_{11} + \tilde{\sigma}_{12})(\tilde{\sigma}_{33} + \tilde{\sigma}_{34}) - 4\tilde{\sigma}_{13}^2]}{(\dot{\sigma}_{11} - \dot{\sigma}_{12})(\dot{\sigma}_{33} - \dot{\sigma}_{34})[(\dot{\sigma}_{11} + \dot{\sigma}_{12})(\dot{\sigma}_{33} + \dot{\sigma}_{34}) - 4\dot{\sigma}_{13}^2]}, \qquad (69)$$

where the tilde indicates ML estimators under H_0 and the dot indicates ML estimators under H_1. The formulas for the terms in the numerator include:

$$\tilde{\sigma}_{11} = \frac{1}{2}(\hat{\sigma}_{11} + \hat{\sigma}_{22}), \qquad (70)$$

$$\tilde{\sigma}_{33} = \frac{1}{2}(\hat{\sigma}_{33} + \hat{\sigma}_{44}), \qquad (71)$$

$$\tilde{\sigma}_{12} = \frac{4(\tilde{\sigma}_{33} - \tilde{\sigma}_{34})\,\bar{\sigma}^2 + (\tilde{\sigma}_{34} - \hat{\sigma}_{34})\,(\tilde{\sigma}_{33} - \tilde{\sigma}_{34})\sigma_{12} - 4(\tilde{\sigma}_{34} - \hat{\sigma}_{34})\,\bar{\sigma}_{34}\tilde{\sigma}_{11}}{(\tilde{\sigma}_{34} - \hat{\sigma}_{34})\,(\tilde{\sigma}_{33} - \tilde{\sigma}_{34}) + 4\bar{\sigma}_{34}\,[(\tilde{\sigma}_{33} - \tilde{\sigma}_{34}) - (\tilde{\sigma}_{34} - \hat{\sigma}_{34})]}, \quad (72)$$

$$\text{where } \bar{\sigma} = \frac{\hat{\sigma}_{13} + \hat{\sigma}_{14} + \hat{\sigma}_{23} + \hat{\sigma}_{24}}{4}, \text{ and} \quad (73)$$

$$\tilde{\sigma}_{34} = \frac{4\,(\tilde{\sigma}_{11} - \tilde{\sigma}_{12})\,\bar{\sigma}^2 + (\tilde{\sigma}_{12} - \hat{\sigma}_{12})(\tilde{\sigma}_{11} - \tilde{\sigma}_{12})\,\hat{\sigma}_{34} - 4(\tilde{\sigma}_{12} - \hat{\sigma}_{12})\tilde{\sigma}_{12}\,\tilde{\sigma}_{33}}{(\tilde{\sigma}_{12} - \hat{\sigma}_{12})(\tilde{\sigma}_{11} - \tilde{\sigma}_{12}) + 4\tilde{\sigma}_{12}\,[(\tilde{\sigma}_{11} - \tilde{\sigma}_{12}) - (\tilde{\sigma}_{12} + \hat{\sigma}_{12})]} \quad (74)$$

Notice that Equations 72 and 74 are simultaneous in two unknowns and have to be solved numerically (at least, the direct solution is unknown). Lord (1957), after finding that his algorithm converged slowly to solution, recommended solving the equations graphically. Given the availability of modern high-speed electronic computer most current users of the test will prefer a rapidly converging numerical analysis technique, e.g., the Newton-Raphson or Fletcher-Powell-Davidon algorithms, to find the values of $\tilde{\sigma}_{12}$ and $\tilde{\sigma}_{34}$. Now we provide the expressions for the denominator:

$$\dot{\sigma}_{jk} = \begin{cases} \hat{\sigma}_{jk}, & \text{if } \hat{r}_{XY} < 1.0, \\ \tilde{\sigma}_{jk}, & \text{otherwise.} \end{cases} \quad (75)$$

Since $\chi = \sqrt{\chi^2}$ is strictly normally distributed, Lord suggests using the positive tail of the normal distribution to obtain the significance level instead of the chi square table. When χ exceeds the tabled value of z at the nominal alpha level, the null hypothesis of a perfect linear relationship between the latent variables X and Y is rejected. When this occurs the ML estimator of r_{XY} under H_1 is given by Lord as

$$\hat{r}_{XY} = \begin{cases} \dfrac{\hat{\sigma}_{13} + \hat{\sigma}_{23} + \hat{\sigma}_{14} + \hat{\sigma}_{24}}{4\sqrt{\hat{\sigma}_{12}\,\hat{\sigma}_{34}}}, & \text{if } \hat{r}_{XY} < 1.0, \\ 1, & \text{otherwise.} \end{cases} \quad (76)$$

Although Lord's method involves considerable computation if undertaken by hand, it could be programmed for solution on an electronic computer. In fact, Jöreskog's analysis of covariance structures program ACOVS, which employs the Fletcher-Powell-Davidon iterative technique to minimize a transform of the likelihood function, can be used to compute Lord's χ^2. The program provides a rapidly converging algorithm which became available only after Lord (1957) had devised his test (Davidon, 1959; Fletcher & Powell, 1963). We will return to this later in the section on Jöreskog's method. Lord's procedure possesses all of the optimal large sample properties of the maximum likelihood technique. However, the degree to which the test is disturbed by departures from parallelism and normality, small sample sizes, etc. is not known. In comparing Lord's (1957) test with other methods reviewed previously, we find that both the Forsyth-Feldt-McNemar and Lord tests evaluate the same null hypothesis that two tests x and y measure the same attribute except for observational errors. While both tests are based on a likelihood ratio principle, Lord's test requires large samples for its justification. Although the Forsyth-Feldt-McNemar test is appropriate for samples of any size, the fact that F_2/k may not follow the central F distribution exactly represents a potential liability of the procedure. The Forsyth-Feldt-McNemar test requires that the reliabilities of x and y be equal; on the other hand Lord's test requires multivariate normality of the observed scores. Finally, we do not know which method is relatively more efficient, that is, how the methods compare in terms of actual frequencies of Type I errors against a specified alpha under the null hypothesis or how they compare in their power to detect true alternative hypotheses. We can conclude that presently there is insufficient evidence for preferring either the Lord or Forsyth-Feldt-McNemar tests.

Lord-Villegas Test

In 1973 Lord demonstrated to psychologists how a statistical procedure developed by Villegas (1964) could be used to test "the hypothesis that two sets of measurements differ only because of errors of measurement and because of differing origins and units of measurement" (Lord, 1973, p. 71). The assumptions of the Lord-Villegas procedure are summarized in Table 2. The reader will note that in addition to the classical assumptions about true and error components, the model requires the errors of measurement from any pair of tests to follow a bivariate normal distribution. However, the errors of measurement for tests x_1 and x_2 may be correlated with those from y_3 and y_4. Although it is not stated by Lord, Kristof (1973) points out that the Lord-Villegas test requires that x_1 be parallel with x_2 and y_3 with y_4. Inspection of Table 2 reveals that the Lord (1957) test differs in assumptions from the Lord-Villegas method primarily in terms of which components are required to be jointly normally distributed and the correlation of errors. Also the Lord-Villegas test does not depend on sample size for its justification.

The null hypothesis tested by the Lord-Villegas procedure is precisely that the two tests x and y are congeneric or scientifically equivalent, i.e., H_0: $r_{XY} = 1.0$. The alternative is that the linear relationship between the true scores is less than perfect.

In order to perform the Lord-Villegas test we must compute three new matrices. Let us define the matrix $\underline{W}$, a within persons matrix, to be

$$\underline{W} = \begin{pmatrix} \Sigma_{x_W}{}^2 & \Sigma_{xy_W} \\ \Sigma_{yx_W} & \Sigma_{y_W}{}^2 \end{pmatrix} , \tag{77}$$

where the elements in $\underline{W}$ are defined as

$$\Sigma_{x_W}{}^2 = \sum_{i=1}^{N} \sum_{j=1}^{2} (x_{ij} - \bar{x}_{i.})^2 , \tag{78}$$

$$\Sigma_{xy_W} = \Sigma_{yx} = \sum_{i=1}^{N} \sum_{j=1}^{2} (x_{ij} - \bar{x}_{i.})(y_{i(j+2)} - \bar{y}_{i.}) , \quad \text{and} \tag{79}$$

$$\Sigma_{y_W}{}^2 = \sum_{i=1}^{N} \sum_{j=3}^{4} (y_{ij} - \bar{y}_{i.})^2 . \tag{80}$$

The reader will recall that all of the symbols are defined in Table 1 except $\bar{x}_{i.}$ and $\bar{y}_{i.}$ which equal $[(x_{i1} + x_{i2})/2]$ and $[(y_{i1} + y_{i2})/2]$, respectively.

Next the among persons sums of squares and cross products matrix $\underline{A}$ is written

$$\underline{A} = \begin{pmatrix} \Sigma_{x_A}{}^2 & \Sigma_{xy_A} \\ \Sigma_{yx_A} & \Sigma_{y_A}{}^2 \end{pmatrix} , \tag{81}$$

where

$$\Sigma_{x_A}{}^2 = 2 \sum_{i=1}^{N} (\bar{x}_{i.} - \bar{x}_{..})^2 , \tag{82}$$

$$\Sigma_{xy_A} = \Sigma_{yx_A} = 2\sum_{i=1}^{N} (\bar{x}_{i.} - \bar{x}_{..})(\bar{y}_{i.} - \bar{y}_{..}), \quad (83)$$

$$\Sigma_{y_A}^2 = 2\sum_{i=1}^{N} (\bar{y}_{i.} - \bar{y}_{..})^2.$$

Now we select the significance level at which we wish to evaluate H_0, say .05. From the tabled values of $\underline{F}$ we find the 1-05 = 95 percentile of the $\underline{F}$ distribution with N and N degrees of freedom. Finally, we evaluate the determinant[2] of matrix $\underline{C}$ of order 2:

$$|\underline{C}| = |\underline{A} - \underline{F}_{95}\underline{W}| . \quad (84)$$

The null hypothesis of equivalent measures is rejected at < .05 significance level if the determinant is positive and if both diagonal terms are also positive, i.e., "H_0 is rejected if and only if the matrix $\underline{C}$ is positive definite" (Lord,1973, p. 71). Lord (1973) explains that the test is slightly conservative in that a true null will be rejected somewhat less often than the value of α would indicate.

The computational simplicity and somewhat less restrictive assumptions of the Lord-Villegas procedure as compared with Lord's (1957) likelihood ratio test would seem to commend it to our use. However, Lord (1973) cautions that the procedure may be very sensitive tocorrelated errors within tests, i.e., $r_{e_1e_2} \neq 0$ and $r_{e_3e_4} \neq 0$. For example, positively correlated errors will tend to increase the elements of $\underline{W}$ and consequently to decrease the probability of rejecting H_0. Although I have not pursued the matter in depth, it appears that correlated errors would affect all of the procedures considered in this review in much the same ways. Thus, the important question concerns which of the methods is the most robust with repsect to violations of the assumption of uncorrelated errors. We observe in closing that the efficiency of this test relative to that of the other procedures has not been researched to date.

Kristof's Test

Kristof's (1973) method for testing if a perfect linear relation exists between the true scores X and Y represents a significant liberalization of the assumptions of the parallelism of x_1 and x_2 and of y_3 and y_4 required by all of the preceding procedures. In fact Table 2 shows that this method makes only three assumptions: 1) that the errors of measurement within test x and within y are uncorrelated, 2) that the errors are not correlated with true scores, and 3) that the errors are multinormally distributed. Kristof's test appears to require fewer restrictive assumptions than any of the other methods.

The null hypothesis is "that two variables have a perfect disattenuated correlation,hence measure the same trait except for errors of measurement. This hypothesis is equivalent to saying, within the adopted model, that true scores of two psychological tests satisfy a perfect linear relation " (Kristof, 1973, p. 101). This may be written as:

$$H_0 : a_1 X + a_2 Y + a_0 = 0 \quad (\text{for } a_1, a_2 \neq 0). \quad (85)$$

The reader should notice that either a_1 or a_2 will be less than zero, i.e., will be negative, under H_0. Equation 82 can be rearranged to clarify the nature of the perfect linear relation:

$$Y = -a_0 - \frac{a_1}{a_2} X. \quad (86)$$

Obviously, if $\mathcal{E}(Y) = \mathcal{E}(X) = 0$ and var(Y) = var(X) = 1.0, then $-a_1/a_2 = r_{XY}$. The alternative hypothesis holds that Equation 85 is nonzero.

In order to derive Kristof's test statistic we define two new variables f_1 and f_2:

$$f_1 = a_1 x_1 + a_2 y_3 = a_1 X_1 + a_2 Y_3 + a_1 e_1 + a_2 e_3, \tag{87}$$

$$f_2 = a_1 x_2 + a_2 y_4 = a_1 X_2 + a_2 Y_4 + a_1 e_1 + a_2 e_4 . \tag{88}$$

(Refer to Equations 10 through 21 for definitions of the variables.) Kristof observed that when H_0 is true, f_1 and f_2 will correlate exactly zero. Thus, H_0 can be reformulated as H_0: $r_{f_1 f_2} = 0$. If $\hat{r}_{f_1 f_2}$ exceeds the critical value of $r_{f_1 f_2}$ for a prespecified α level, then H_0 is rejected. The test is conservative according to Kristof (1973) so that "if rejection of H_0 occurs, then the true corresponding level $\bar{\alpha}$ will not exceed α, $\bar{\alpha} \leq \alpha$" (p. 108). In order to test H_0, we compute the minimum possible value of $f_{f_1 f_2}$ given the data subject to the restriction that a_1, $a_2 \neq 0$. Letting $\hat{r}_{min}$ be the minimum value of $\hat{r}_{f_1 f_2}$, compute

$$\underline{t} = \frac{\hat{r}_{min}\sqrt{N-2}}{\sqrt{1 - \hat{r}_{min}^2}} . \tag{89}$$

If the sample value of $\underline{t}$ exceeds the tabled value of $\underline{t}$ for N-2 $\underline{df}$, then we reject the hypothesis of equivalence of x and y. A one-tailed test is performed because of the assymetry of the alternative hypothesis, H_1: $r_{XY} < 1.0$.

Kristof (1973) describes several tests which are based on different assumptions concerning the parameters of the error distributions. However, we shall develop only the least restricted model. As a first step in computing $\hat{r}_{min}$ we rearrange the rows and columns in $\hat{\Sigma}_{xy}$ to form $\hat{\Sigma}$

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{11} & \vdots & \hat{\Sigma}_{12} \\ \cdots & \vdots & \cdots \\ \hat{\Sigma}'_{12} & \vdots & \hat{\Sigma}_{22} \end{pmatrix} = \begin{pmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{13} & \vdots & \hat{\sigma}_{12} & \hat{\sigma}_{14} \\ \hat{\sigma}_{31} & \hat{\sigma}_{33} & \vdots & \hat{\sigma}_{32} & \hat{\sigma}_{34} \\ \cdots & \cdots & \vdots & \cdots & \cdots \\ \hat{\sigma}_{21} & \hat{\sigma}_{23} & \vdots & \hat{\sigma}_{22} & \hat{\sigma}_{24} \\ \hat{\sigma}_{41} & \hat{\sigma}_{43} & \vdots & \hat{\sigma}_{42} & \hat{\sigma}_{44} \end{pmatrix} . \tag{90}$$

Next an eigen decomposition of $\hat{\Sigma}_{12}$ is performed to yield orthogonal matrices $\underline{P}$ and $\underline{S}$ of order 2:

$$\hat{\Sigma}_{12} = \underline{P}\,\underline{S}\,\underline{P}'$$

In the next step new matrices $\underline{Q}$ and $\underline{U}$ are found as follows:

$$\underline{Q} = \underline{S}^{-\frac{1}{2}}\,\underline{P}'\,\hat{\Sigma}_{11}\,\underline{P}\,\underline{S}^{-\frac{1}{2}} = (q_{jk}) \text{ for } j = 1,2;\ k = 1,2 \text{, and} \tag{92}$$

$$\underline{U} = \underline{S}^{-\frac{1}{2}}\,\underline{P}'\,\hat{\Sigma}_{22}\,\underline{P}\,\underline{S}^{-\frac{1}{2}} = (u_{jk}) \text{ for } j = 1,2;\ k = 1, 2 . \tag{93}$$

Finally, we solve the quartic trigonometric equation

$$h_4 \cot^4 v + h_3 \cot^3 v + h_2 \cot^2 v + h_1 \cot v + h_0 = 0, \quad (94)$$

where

$$h_4 = q_{11} u_{12} + q_{12} u_{11}, \quad (95)$$

$$h_3 = q_{11}(u_{22} - u_{11}) + u_{11}(q_{22} - q_{11}) + 4 q_{12} u_{12}, \quad (96)$$

$$h_2 = 3 [q_{12} (u_{22} - u_{11}) + u_{12} (q_{22} - q_{11})], \quad (97)$$

$$h_1 = q_{22} (u_{22} - u_{11})(q_{22} - q_{11}) - 4 q_{12} u_{12}, \quad (98)$$

$$h_0 = -q_{12} u_{22} - q_{22} u_{12}. \quad (99)$$

There will be four solutions of this equation, two of which must be real. From the largest root of the quartic we find $\hat{r}_{min}$ which is then used in the formula for $\underline{t}$ (Equation 89) to evaluate the null hypothesis.

Although Equations 88 through 96 may appear formidable, they are easily and quickly solved by standard computer programs. Most computer installations will have a program for solving 4th degree polynomial equations in a single variable, say d. To solve Equation 94, let d = cot v and use the program to obtain the largest real root, say s, of

$$h_4 d^4 + h_3 d^3 + h_2 d^2 + h_1 d + h_0 = 0. \quad (100)$$

Then the corresponding root s' of Equation 94, can be obtained by using the inverse trignometric function to solve : s' = arccot s. It appears that Kristof's (1973) method represents an efficient procedure for testing if $r_{xy} = 1.0$ under very liberal assumptions. Although the test is conservative, it is valid in small sample applications

Thus, Kristof's (1973) procedure possesses some real advantages over those methods already reviewed. To summarize, the test does not require the parallelism of x_1 and x_2 and of y_3 and y_4 and permits some between-test error correlations. Large samples are not required to justify the validity of the test. Widely-available standard computer programs for solving polynominal equations can be used to compute the necessary test statistic quickly and inexpensively. However, analytical and empirical studies have not been performed so that the relative efficiency of the test is unknown. It would be especially important to researchers who must choose one of these methods to know the comparative frequencies of Type I and II errors for the Kristof, Lord, Lord-Villegas, and Forsyth-Feldt-McNemar tests when the split tests satisfy only congeneric assumptions. In other words, does use of Kristof's procedures convey a tangible benefit to the user when x_1, x_2 and y_3, y_4 are not parallel? While it seems reasonable to anticipate an affirmative answer, we simply do not have any data on which to compare the methods. Empirically or analytically derived reasons for preferring Kristof's test to the others will have to come from future research.

Joreskog's Test

Jöreskog's maximum likelihood method for estimating the parameters of and testing hypotheses about covariance structures provides a very flexible approach for investigating the

equivalence of measures. As can be gleaned from Table 2 the method is based upon the large sample properties of maximum likelihood estimators and likelihood ratio tests subject to classical test theory assumptions and the multinormality of the observation vector (x_1, x_2, y_3, y_4). Parenthetically, we mention that the general ACOVS model can be defined so as to relax the classical assumption of uncorrelated errors. However, more replicate measures on x and y may be required in order to identify the model. It is obvious from Table 2 that Jöreskog's method compares favorably with the procedure of Kristof (1973) discussed in the last section. However, as will be shown below Jöreskog's technique allows greater flexibility, because it allows one to test a variety of restrictions and hypotheses.

For the purposes of this review we are interested specifically in testing two different null hypotheses within the framework of Jöreskog's ACOVS model. The first is that $r_{XY} = 1.0$ which we now write as H_E (for equivalence). The second null is that for each variable the half tests meet the equality of units of measurement and standard errors assumptions for parallel tests. Thus H_P symbolizes the null hypothesis that x_1 and x_2 are parallel and that y_3 and y_4 are also parallel. Of course, Jöreskog's method is completely general so that an interested investigator could test assumptions of the tau equivalence, of the parallelism of four observed variables, i.e., Wilks (1946) test, etc. Although in this section equality constraints on the means or origins of the measures will not be considered, the reader should appreciate that Jöreskog's general model readily permits tests about the structuring of the means (as was illustrated in a previous section). There is typically little interest in differences in means between tests, so this issue will not be pursued here. After considering the compututational formulas, we shall describe how two alternative tests of H_E and H_p may be formulated and evaluated.

Jöreskog's (1970) general method for analyzing covariance structures assumes that the population covariance matrix Σ_{xy} (Equation 39) has the form given in Equation (18) which is repr oduced here for the reader's convenience:

$$\Sigma_{xy} = \underline{B}\, \Sigma_{XY}\, \underline{B}' + \Sigma_{ee} \quad . \tag{18}$$

A covariance matrix of this structure is produced when the observed variables are structured as Equation 22 (reproduced here):

$$\underline{v} = \underline{m} + \underline{B}\, \underline{t} + \underline{e} \quad . \tag{22}$$

Three kinds of parameters may be contained in the parameter matrices $\underline{B}$, Σ_{XY}, and Σ_{ee}: (i) <u>fixed parameters</u> that are assigned a priori values, (ii) <u>constrained parameters</u> that are unknown, but equal to one or more other parameters, and (iii) <u>free parameters</u> that are unknown and unconstrained.

The problem is to find estimates of the constrained and free parameters which maximize the likelihood of the sample values given a model of the form of Equation 18. For most applications simple analytic solutions do not exist (Wilks(1949) test is one in which they do), so Jöreskog (1970) uses the numerical method of Davidon (1959) and Fletcher and Powell (1963) to maximize the likelihood function. Compared with variants of the Newton-Raphson technique, this is an efficient procedure which makes use of the derivatives of the likelihood function and the inverse of the information matrix. Actually, Joreskog (1970) finds it more convenient to minimize a function O, which is equivalent to maximizing the logarithm of the likelihood function L:

$$O = \log \left| \tilde{\Sigma}_{xy} \right| + \mathrm{tr} \left(\hat{\Sigma}_{xy}\, \tilde{\Sigma}_{xy}^{-1} \right) - \log \left| \hat{\Sigma}_{xy} \right| - J, \tag{101}$$

where $\tilde{\Sigma}_{xy}$ contains the maximum likelihood estimates of $\tilde{\sigma}_{jk}$ estimated under the model

specified by Equation 18 and J is the number of observed variables. O is a function of the independent elements in $\underline{B}$, Σ_{XY}, and Σ_{ee}. In large samples, N-1 times the minimum value of O is distributed as chi square and may be used to test the goodness of fit of the model. In addition approximate standard errors may be obtained for each estimated parameter from the inverse of the information matrix computed at the minimum of O.

Hypotheses are tested in this approach by the likelihood ratio technique. The ACOVS program (Jöreskog, Gruvaeus, & van Thillo, 1970) computes a chi square value for each specified model against the most general alternative that Σ_{xy} is any positive definite matrix:

$$\chi^2 = -2 \ln \frac{L(\Sigma_R)}{L(\Sigma_F)} , \tag{102}$$

where $L(\Sigma_R)$ represents the likelihood under a given specification of fixed, free and constrained parameters (Restricted model), and $L(\Sigma_F)$ is the likelihood under the assumption that Σ_{xy} is any positive definite matrix (Full model). According to Jöreskog (1970) it is possible to test any given model, say M_{R_1}, against a more general alternative, say M_{R_2}, by estimating and testing each one separately (against the most general alternative that Σ_{xy} is any p.d. matrix) and comparing their χ^2 goodness of fit values. The difference in chi square values is asymptotically chi square distributed with degrees of freedom equal to the corresponding difference in degrees of freedom between the two models:

$$\chi_D^2 = \chi_{R_1}^2 - \chi_{R_2}^2 = N(O_{R_1} - O_{R_2}) \tag{103}$$

with

$$df_D = df_{R_2} - df_{R_1} \tag{104}$$

In general the number of degrees of freedom on which any chi square test is based equals the difference in the number of parameters estimated under the full and restricted models.

With this introduction to Jöreskog's method we can now explicate the hypotheses (models) of interest in investigations seeking to determine the equivalence of measures. Following Jöreskog (1971) we suggest four models which could be tested:

M_1 : $b_1 = b_2$, $b_3 = b_4$, $\sigma_{e_1}^2 = \sigma_{e_2}^2$, $\sigma_{e_3}^2 = \sigma_{e_4}$, $r_{XY} = 1.0$

M_2 : $b_1 = b_2$, $b_3 = b_4$, $\sigma_{e_1}^2 = \sigma_{e_2}^2$, $\sigma_{e_3} = \sigma_{e_4}^2$

M_3 : $r_{XY} = 1.0$

M_4 : Σ_{xy} is any p.d. matrix of rank 2 with the elements of $\underline{B}$, Σ_{xy}, and Σ_{ee} all free.

Each of these four models is tested against the most general model:

M_5: Σ_{xy} is any p.d. matrix.

This series of tests is illustrated in the upper portion of Table 3 where the numbers of parameters and degrees of freedom are indicated. To test the hypothesis that the two tests x and y are equivalent (H_E: $r_{xy} = 1.0$) we could consider the goodness of fit

---

Insert Table 3 about here

---

of either models 1 or 3. H_P, the null hypothesis that x_1, x_2 and y_3, y_4 are parallel, could be tested by model 2. However, Joreskog (1974) maintains that "the value of χ^2 should be interpreted very cautiously." He suggests that it is more informative to test the reasonableness of any restriction by fitting two different models, one of which contains the restriction, the other of which does not. "The differences between χ^2 values matter rather than the χ^2 values themselves " (Jöreskog, 1974).

In the lower portion of Table 3 are presented four model comparisons which yield tests of H_E and H_P. The differences in chi square values for the indicated models yield tests of the following null hypotheses:

M_1 v. M_2: Given x_1, x_2 and y_3, y_4 are parallel, test if x and y are congeneric.

M_3 v. M_4: Given x_1, x_2 and y_3, y_4 are congeneric, test if x and y are congeneric.

M_3 v. M_1: Given x and y are congeneric, test if x_1, x_2 and y_3, y_4 are parallel.

M_4 v. M_2: Given x and y are not congeneric, test if x_1, x_2 and y_3, y_4 are parallel.

(Of course other possibilities exist, e.g. tau equivalence, and these can be tested easily by the ACOVS program.)

We observe that the test of the M_1 v. M_2 comparison is identical to Lord's (1957) test. They differ only in computing algorithms, and as was suggested earlier, the Fletcher-Powell-Lavidon method used by Joreskog might prove satisfactory in solving Equations 72 and 74. The comparison of models 3 and 4 is comparable to Kristof's (1973) Case III which was presented in the prior section. Although the underlying assumptions and null and alternative hypotheses are (roughly) the same, Kristof's and Joreskog's test statistics differ considerably. Unfortunately, no one has undertaken an analytic comparison of the two methods so that potential users of the procedures do not know which of the two methods performs best with data with similar characteristics. For example, is Kristof's method to be preferred when N is less that 100? With large samples is Joreskog's test more powerful? Which technique is more sensitive to departures from the normality assumptions? Educational and psychological researchers who desire to investigate the equivalence of a set of tests require answers to these kinds of questions in order to select the procedure with optimal characteristics for their problems.

Joreskog's method affords the capability of testing which test model is appropriate for x_1 and x_2 and of y_3 and y_4. Identical tests of parallelism can be constructed using

the methods of Wilks (1946), Votaw (1948), and Jöreskog (1970, 1971). These all produce likelihood ratio tests, but they differ in computational methods. The ease and flexibility of Jöreskog's procedure would seem to recommend it for testing assumptions about test score models, e.g., whether a set of measurements conforms to the assumptions of congeneric, essentially tau equivalent, or parallel tests.

In concluding this description of the ACOVS method we point out that it will yield a ML estimate of r_{XY} when H_E is not tenable and that it easily accommodates the analysis of several sets of congeneric tests (e.g., x, y, z) each of which has several replications (e.g., x_1, x_2, x_3, y_4, y_5, z_1, z_2, z_3). The null hypothesis of interest in this situation is H_0: $r_{XY} = r_{XZ} = r_{YZ} = 1.0$. Finally, it is interesting to note that ACOVS can be used to test the hypothesis that the correlation corrected for attenuation equals 1.0 in situations where replicate measurements on x and y are not available if the reliabilities of x and y are known (cf. Equation 3). The variances of the errors of measurement are determined from the reliabilities according to the psychometric identity $\delta_e^2 = (1 - r_{xx})\delta_x^2$ (Gulliksen, 1950). In the restricted model the values of δe_x, δe_y, and r_{XY} are fixed, while b_1 and b_2 are free. Thus, the restricted model estimates 2 parameters. (See Figure 3.) The alternative hypothesis is based on $2(2 + 1) / 2 = 3$ parameters. The difference between the number of parameters in the full and restricted models yields one overidentifying restriction on which to

---

Insert Figure 3 about here

---

test the goodness of fit of the restricted model (H_0: $r_{XY} = 1.0$). The interested reader is referred to Jöreskog (1970, 1971, 1974) and Jöreskog, et al. (1970) for more details.

Numerical Example

Although only three of the methods for determining equivalence have been systematically compared (the McNemar, Forsyth-Feldt-McNemar, and Forsyth-Feldt tests), all the procedures which we have reviewed have been or can be applied to the data set first analyzed by Lord (1957). The following sample covariance and correlation matrices are based on a sample of 649 individuals who were administered two (supposedly) parallel forms (x_1 and x_2) of an unspeeded 15-item vocabulary test (x) and two (supposedly) parallel forms (y_1 and y_2) of a highly speeded 75-item vocabulary test (y):

$$\hat{\Sigma}_{xy} = \begin{pmatrix} 86.3979 & & & \\ 57.7751 & 86.2632 & & \\ 56.8651 & 59.3177 & 97.2850 & \\ 58.8986 & 59.6683 & 73.8201 & 97.8192 \end{pmatrix}$$

$$\hat{R}_{xy} = \begin{pmatrix} 1.0000 & & & \\ .6692 & 1.0000 & & \\ .6203 & .6475 & 1.0000 & \\ .6407 & .6496 & .7567 & 1.0000 \end{pmatrix}$$

As Lord (1957) states it, "The question here is whether or not the speeded and unspeeded vocabulary tests actually measure the same ability" (pp. 210-211).

In Table 4 are reported the results of analyzing Lord's data by each of the methods reviewed here. The values of the test statistics, degrees of freedom, significance level

---

Insert Table 4 about here

---

and estimate of the disattenuated correlation (if available) are given. All of the methods reject the hypothesis of equivalence at $p < .001$. Given data with these characteristics and an N as large as 649 there do not appear to be any differences in the performances of the tests. The estimate of the correlation between true scores X and Y is approximately .9 given estimates of $r_{xx} = .80$ and $r_{yy} = .86$ according to three of the methods. It is worth mentioning that for all of the methods the null hypothesis would have been rejected at $p < .05$ if the N had been only 101. Finally, we note that as would be expected the chi squares for Lord's (1957) test and Jöreskog's $M_1v_1M_2$ comparison differ by only .2.

Discussion

This completes our review of the eight methods for determining the equivalence of measures which have been found in the statistical and measurement journals of psychology and education. Other procedures for estimating linear functional relationships have been devised by mathematical statisticians and econometricians. However, their major concern has been with the estimation rather than the testing of such models, and so these techniques have not been covered here. See Kendall and Stuart (1961), Madansky (1959), and Moran (1971) for reviews. Our survey of the methods due to Wilks (1948), McNemar (1958), Forsyth and Feldt (1969, 1970), Lord (1957, 1973), Kristof (1973), and Jöreskog (1971) leads to the following conclusions.

First, we have discovered that some of the methods, although utilizing different computational algorithms, yield identical statistical tests of the null hypothesis. Jöreskog's maximum likelihood method for estimating the general ACOVS model and the resulting chi square goodness of fit test subsumes the likelihood ratio tests of Wilks (1948) and Lord (1957). Given the greater flexibility of the ACOVS approach and the ready availability of associated computer programs,[3] we think that most researchers will perfer Jöreskog's method to those of Wilks (1946) and Lord (1957).

Our second conclusion is that Forsyth and Feldt's (1970) modification of McNemar's (1958) test gives better performance than either the original McNemar test or the Forsyth and Feldt (1969) approximate normal curve procedure. Monte Carlo studies (Forsyth & Feldt, 1969, 1970) were reviewed that demonstrated the greater efficiency of the Forsyth-Feldt-McNemar test. Thus, the interested researcher has effectively four alternatives from which to choose a statistical procedure for evaluating the hypothesis that two tests measure the same latent trait: the Jöreskog, Forsyth-Feldt-McNemar, Lord-Villegas, and Kristof methods.

The third point to be made in this summary is that these four methods differ substantially in the assumptions they require and the properties they possess. Two of the tests, Forsyth-Feldt-McNemar and Lord-Villegas, require that the half tests be parallel. The tests of Jöreskog and Kristof are applicable in more situations since they assume only that x_1, x_2, and y_3, y_4 are congeneric. While Jöreskog's ACOVS is the most flexible method and is widely available as a computer program, it requires large samples for the justification of its asymptotic properties. The quality of the performance of maximum likelihood estimation (MLE) in small samples is always suspect. As Wonnacott and Wonnacott (1972) conclude, "although MLE has the many attractive large sample properties..., in small sample estimation it should be used with considerable caution" (p. 460). The validity of the other three tests does not depend upon sample size. However, the simulation studies of Forsyth and Feldt (1970) showed that the Forsyth-Feldt-McNemar test was unsafe in the sense that the empirical alpha level exceeds that of the nominal alpha. On the other hand, according to the authors of the procedures the Lord-Villegas and Kristof tests are conservative, i.e., the actual alpha is less than the nominal level. What effects these differences would have on the comparative rates of Type I errors among the methods is unknown.

The final conclusion based on our review of the four methods can be simply put: there is no information available for choosing among the tests on the basis of relative efficiency. We do not know which method maintains the most effective control of Type I errors for fixed sample size. Also there have been no studies which compared the power of the four procedures to detect true alternative hypotheses. Thus, investigators who wish to use the most efficient test lack the data necessary to make the best choice. In closing we point out that the differences in sensitivity to violations of assumptions (e.g., nonnormality or correlated errors) among the four methods have not been investigated. There is insufficient information available for choosing the most robust statistical test.

References

Aigner, D., & Goldberger, A.S. (Eds.) Latent variables in socio-economic models. Amsterdam: North-Holland, 1976.

American Psychological Association. Manual for educational and psychological tests. Washington, D.C.: Author, 1966.

Blalock, H.M. Multiple indicators and the causal approach to measurement error. American Journal of Sociology, 1969, 75, 264-272.

Block, J. The equivalence of measures and the correction for attenuation. Psychological Bulletin, 1963, 60, 152-156.

Box, G.E.P., & Muller, M.E. A note on the generation of random normal deviates. Annals of Mathematical Statistics, 1958, 29, 610-611.

Campbell, D.T., & Fiske, D.W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Cronbach, L.J. Test validation. In R.L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Cronbach, L.J., & Meehl, P.E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 456-473.

Cureton, E.E. On certain estimated correlation functions and their standard errors. Journal of Experimental Education, 1936, 4, 252-264.

Davidon, W.C. Variable metric method for minimization. A.E.C. Research and Development Report, ANL-5990 (Rev.), 1959.

Fletcher, R., & Powell, M.J.D. A rapidly converging descent method for minimization. The Computer Journal, 1963, 6, 163-168.

Forsyth, R.A., & Feldt, L.S. An investigation of empirical sampling distributions of correlation coefficients corrected for attenuation. Educational and Psychological Measurement, 1969, 29, 61-72.

Forsyth, R.A., & Feldt, L.S. Some theoretical and empirical results related to McNemar's test that the population correlation coefficient corrected for attenuation equals 1.0. American Educational Research Journal, 1970, 7, 197-207.

Goldberger, A.S., & Duncan, O.D. Structural equation models in the social sciences. New York: Seminar Press, 1973.

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Gulliksen, H. Methods for determining equivalence of measures. Psychological Bulletin, 1968, 70, 534-544.

Hammersley, J.M., & Handscomb, D.C. Monte Carlo methods. Methuen, 1964.

Hanushek, E.A., & Jackson, J.E. Statistical methods for social scientists. New York: Academic Press, 1977.

Hsu, T.C., & Sebatane, E.M. The effect of differences in covariate means among treatment groups upon the F-test of the analysis of covariance. Paper presented at the Annual Meeting of the American Educational Research Association, New York, N. Y., April, 1977.

Isaac, P.D. Linear regression, structural relations, and measurement error. Psychological Bulletin, 1970, 74, 213-218.

Johnston, J. Econometric methods. New York: McGraw-Hill, 1963.

Jöreskog, K.G. A general method for analysis of covariance structures. Biometrika, 1970, 57, 239-251.

Jöreskog, K.G. Statistical analysis of sets of congeneric tests. Psychometrika, 1971, 36, 109-133.

Jöreskog, K.G. Analyzing psychological data by structural analysis of covariance matrices. In D.H. Krantz, R.C. Atkinson, R.D. Luce, & P. Suppes (Eds.), Contemporary developments in mathematical psychology (Vol. II). San Francisco: Freeman, 1974.

Jöreskog, K.G., & Goldberger, A.S. Estimation of a model with multiple indicators and multiple causes of a single latent variable. Journal of the American Statistical Association, 1975, 70, 631-639.

Jöreskog, K.G., Gruvaeus, G.T., & van Thillo, M. ACOVS-A general computer program for analysis of covariance structures (RB 70-15). Princeton, N.J.: Educational Testing Service, 1970.

Kelley, T.L. Fundamentals of statistics. Cambridge: Harvard University Press, 1947.

Kendall, M.G., & Stuart, A. The advanced theory of statistics, Vol. 2: Inference and relationship. London: Charles Griffin, 1961.

Kristof, W. Testing a linear relation between true scores of two measures. Psychometrika, 1973, 38, 101-111.

Lord, F.M. A significance test for the hypothesis that two variables measure the same trait except for errors of measurement. Psychometrika, 1957, 22, 207-220.

Lord, F.M. Testing if two measuring procedures measure the same dimension. Psychological Bulletin, 1973, 79, 71-72.

Lord, F.M., & Novick, M.R. Statistical theories of mental test scores (with contributions by Allan Birnbaum). Reading, Mass.: Addison-Wesley, 1968.

Madansky, A. The fitting of straight lines when both variables are subject to error. Journal of the American Statistical Association, 1959, 54, 173-205.

McNemar, Q. Attenuation and interaction. Psychometrika, 1958, 23, 259-265.

Mendenhall, W., & Scheaffer, R.L. Mathematical statistics with applications. North Scituate, Mass.: Duxbury Press, 1973.

Moran, P.A.P. Estimating structural and functional relationships. Journal of Multivariate Analysis, 1971, 1, 232-255.

Muller, M.E. A comparison of methods for generating normal deviates on digital computers. Journal of the Association for Computing Machinery, 1959, 6, 376-383.

Nesselroade, J.R., & Reese, H.W. Life-span developmental psychology: Methodological issues. New York: Academic Press, 1973.

Spearman, C. The proof and measurement of association between two things. American Journal of Psychology, 1904, 15, 72-101. (a)

Spearman, C. General intelligence, objectively determined and measured. American Journal of Psychology, 1904, 15, 201-293. (b)

Spearman, C. Demonstration and formulae for true measurement of correlation. American Journal of Psychology, 1907, 18, 161-169.

Spearman, C. Coefficient of correlation calculated from faulty data. British Journal of Psychology, 1910, 3, 271-295.

Villegas, C. Confidence region for a linear relation. Annals of Mathematical Statistics, 1964, 35, 780-788.

Votaw, D.F., Jr. Testing compound symmetry in a normal multivariate distribution. Annals of Mathematical Statistics, 1948, 19, 447-473.

Werts, C.E., & Linn, R.L. Corrections for attenuation. Educational Psychological Measurement, 1972, 32, 117-127.

Werts, C.E., Linn, R.L., & Jöreskog, K.G. Another perspective on "Linear regression, structural relations, and measurement error." Educational and Psychological Measurement, 1973, 33, 327-332.

Wilks, S.S. Sample criteria for testing equality of means, equality of variance and equality of covariances in a normal multivariate distribution. Annals of Mathematical Statistics, 1946, 17, 257-281.

Wonnacott, T.H., & Wonnacott, R.J. Introductory statistics (2nd ed.). New York: Wiley, 1972.

FOOTNOTES

[1]In this quotation and all others cited, symbols have been changed to conform to the notational conventions used in this paper.

[2]The determinant of a 2x2 matrix is equal to the product of the diagonal elements minus the product of the off-diagonal elements.

[3]The most recent version of Jöreskog's program for the analysis of covariance structures COnfirmatory Factor Analysis with Model Modification (COFAMM) is marketed through International Educational Resources, Inc., Box A 3650, Chicago, Illinois 60690.

Table 1

Score Schema[a]

	Test					
	x			y		
	Replication			Replication		
Individual	(1)	(2)	Sum	(3)	(4)	Sum
1	x_{11}	x_{12}	$x_{1\cdot}$	y_{13}	y_{14}	$y_{1\cdot}$
2	x_{21}	x_{22}	$x_{2\cdot}$	y_{23}	y_{24}	$y_{2\cdot}$
.						
.						
.						
i	x_{i1}	x_{i2}	$x_{i\cdot}$	y_{i3}	y_{i4}	$y_{i\cdot}$
.						
.						
.						
N	x_{N1}	x_{N2}	$x_{N\cdot}$	y_{N3}	y_{N4}	$y_{N\cdot}$
Mean	$\bar{x}_{\cdot 1}$	$\bar{x}_{\cdot 2}$	$\bar{x}_{\cdot\cdot}$	$\bar{y}_{\cdot 3}$	$\bar{y}_{\cdot 4}$	$y_{\cdot\cdot}$

[a]Adapted from McNemar (1958, p. 259).

Table 2

Comparison of Assumptions and Hypotheses of Eight Methods for Determining Equivalence

	Test[a]							
Assumption	Wilks-Votaw	McNemar	Forsyth-Feldt-McNemar	Forsyth-Feldt	Lord	Lord-Villegas	Kristof	Jöreskog
Large sample test	Y	N	N	Y	Y	N	N	Y
x_1 and x_2 same origins	N/H	Y	Y	Y	Y	Y	N	N/H
x_1 and x_2 same units of measurement	H	Y	Y	Y	Y	Y	N	N/H
x_1 and x_2 same std error of measurement	H	Y	Y	Y	Y	Y	N	N/H
y_3 and y_4 same origins	H	Y	Y	Y	Y	Y	N	N/H
y_3 and y_4 same units of measurement	H	Y	Y	Y	Y	Y	N	N/H
y_3 and y_4 same std error of measurement	H	Y	Y	Y	Y	Y	N	N/H
x_1, x_2, y_4 same origins	H	Y	Y	N	N	N	N	N/H
x_1, x_2, y_3 and y_4 same units of measurement	H	Y	Y	N	N	N	N	N/H
x_1, x_2, y_3 and y_4 same std error of measurement	H	Y	N	N	N	N	N	N/H
x and y same reliabilities	H	Y	N	N	N	N	N	N/H
$r_{XY} = 1.0$	H	H	H	H	H	H	H	H
$\mathcal{E}(e_1e_2) = 0$, $\mathcal{E}(e_3e_4) = 0$	Y	Y	Y	Y	Y	Y	Y	Y/H
$\mathcal{E}(e_jX) = \mathcal{E}(e_jY) = 0$	Y	Y	Y	Y	Y	Y	Y	Y/H
$\mathcal{E}(e_1e_3) = \mathcal{E}(e_1e_4) = \mathcal{E}(e_2e_3) = \mathcal{E}(e_2e_4) = 0$	Y	Y	Y	Y	Y	N	N	Y/H
$\mathcal{E}(e_1) = \mathcal{E}(e_2) = \mathcal{E}(e_3) = \mathcal{E}(e_4) = 0$	Y	Y	Y	Y	Y	Y	N	Y
x_1, x_2, y_3 and y_4 multivariate normal	Y	N	N	N	Y	N	N	Y
e_1, e_2, e_3 and e_4 normally distributed	N	Y	Y	N	N	Y	Y	N
e_1e_3, e_1e_4, e_2e_3 and e_2e_4 bivariate normal	N	N	N	N	N	Y	Y	N
e_1e_3 and e_1e_4 same joint distributions	N	N	N	N	N	Y	N	N
e_2e_3 and e_2e_4 same joint distributions	N	N	N	N	N	Y	N	N
Easily generalized to three or more tests (x, y, z,...)	Y	Y	N	N	N	N	N	Y
x and y bivariate normal distribution	N	N	N	Y	Y	N	N	N

[a]The letters make the following designations: Y - Yes, the assumption is required; N - No, the assumption is not required; H - Hypothesis, the assumption is tested as the null hypothesis.

Table 3

Tests of Equivalence Using ACOVS

Restricted Model	Number of Parameters	Full Model	Number of Parameters	df for χ^2_D
M_1	4	M_5	10	6
M_2	5	M_5	10	5
M_3	8	M_5	10	2
M_4	9	M_5	10	1
M_5	10	--Not tested--		
M_1	4	M_2	5	1
M_3	8	M_4	9	1
M_3	2	M_1	6	4
M_4	1	M_2	5	4

Table 4

Results of Tests of Equivalence on Lord's (1957) Numerical Example

Test[a]	Test Statistic or Critical Value	df	p	r_{XY}
Wilks-Votaw*	$\chi^2_{vc} = 60.78$	8	<.001	
McNemar	$F = 10.423$	648,1297	<.001	
	$z = 6.19$		<.001	
Forsyth-Feldt-McNemar*	$F = 10.423$	648,1297	<.001	
Forsyth-Feldt*	$r_{crit} = .9059$		<.001	.899
Lord	$\chi^2 = 35.30$	1	<.001	.899
	$\chi = z = 5.94$		<.001	
Lord-Villegas	$\lvert M \rvert = 9.92E8$	649,649	<.001	
Kristof	$t = 4.76$	647	<.001	
Jöreskog				
M_1 v. M_2	$\chi^2 = 35.40$	1	<.001	.899
M_3 v. M_4	$\chi^2 = 35.51$	1	<.001	
M_3 v. M_1	$\chi^2 = 1.12$	4	ns	
M_4 v. M_2	$\chi^2 = 1.23$	4	ns	

[a]An asterisk marks those tests which were performed by this author. All of the others were reported by the developer of the test in the article cited in the text.

Table 5

Comparison of Test Score Models[a]

Test Score Model	Propensity Distributions: First Two Moments	Propensity Distributions: Higher Moments	Experimental Independence	Linear Experimental Independence	True Scores	Error Variances	Observed Means	Intercorrelations	Validities
Equivalent	Equal	Equal	Yes	Yes	Equal	Equal	Equal	Equal	Equal
Parallel	Equal	Unequal	No	Yes	Equal	Equal	Equal	Equal	Equal
T- equivalent	Unequal	Unequal	No	Yes	Equal	Unequal	Equal	Unequal	Unequal
Essentially T- equivalent	Unequal	Unequal	No	Yes	Unequal[b]	Unequal	Unequal	Unequal	Unequal
Congeneric	Unequal	Unequal	No	Yes	Unequal[c]	Unequal	Unequal	Unequal	Unequal

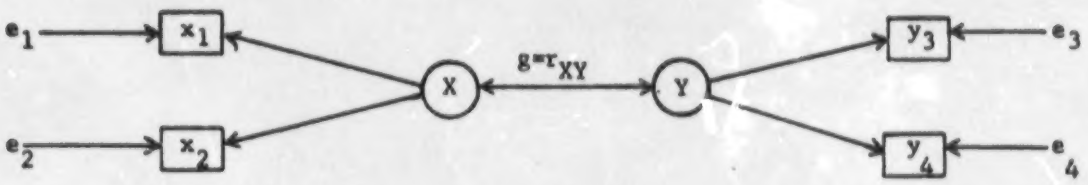
[a] See Lord and Novick (1968, Ch. 2) for more information

[b] True scores may differ only by an additive constant

[c] True scores may differ only by an additive constant and a scaling factor

Figure 1

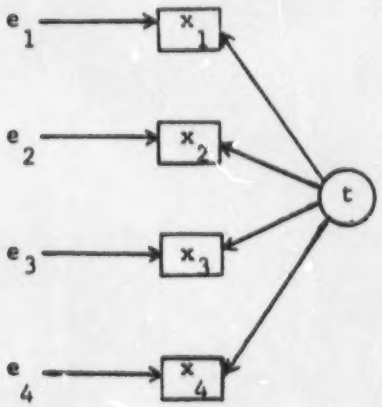
Path Model with Two Sets of Congeneric Tests[a]



[a]Adapted from Jöreskog (1974), Figure 3.

Figure 2

Path Model with Four Congeneric Tests[a]



[a]Adapted from Jöreskog (1974), Figure 1.

Figure 3

Path Model for Correlation Corrected for Attenuation